

Predicting Document Access in Large, Multimedia Repositories

Mimi M. Recker
James Pitkow

College of Computing
Graphics, Visualization & Usability Center
Georgia Institute of Technology
Atlanta, GA 30332-0280 U.S.A.
+1 404.894.9218
mimi@cc.gatech.edu
pitkow@cc.gatech.edu

September 5, 1994

Abstract

Network-accessible multimedia databases, repositories, and libraries are proliferating at a rapid rate. A crucial problem for these repositories remains timely and appropriate document access. In this paper, we borrow a model from psychological research on human memory, which has long studied retrieval of memory items based on frequency and recency rates of past item occurrences. Specifically, the model uses frequency and recency rates of prior document accesses to predict future document requests. The model is illustrated by analyzing the log file of document accesses to the Georgia Institute of Technology World-Wide Web (WWW) database, a large multimedia repository exhibiting high access rates. Results show that the model predicts document access rates with a reliable degree of accuracy. We describe extensions to the basic approach that combine the recency and frequency analyses, and incorporate repository structure and document type. These results have implications for the formulation of descriptive user models of information access in large repositories. In addition, we sketch applications in the areas of design of information systems and interfaces, and their document caching algorithms.

Introduction

Network-accessible multimedia databases, repositories, and digital libraries are proliferating at a rapid rate. These repositories offer the potential for many users, in geographically diverse locations, to access and browse vast quantities of interconnected information – all in synchrony [Nielsen, 1990]. The recent availability of such resources provides new opportunities for analyzing and understanding users as they access and seek information. In particular, the analysis of user patterns allows the formulation of descriptive user models of information access in large, distributed information spaces. Additionally, these models can contribute to the design of information systems, their user interfaces, and document caching strategies.

To approach this problem, we considered other domains where items are retrieved at a high rate from a large available pool. Two potentially related domains are human memory (where many items are recalled on a daily basis) and libraries (where many books are borrowed on a daily basis). Both domains are the locus of extensive research in the fields of psychology and library science.

In psychology, there exists a long tradition of research on human memory. This research has typically focused on people’s recall rates for particular items. Results suggest that the strongest predictors of recall accuracy are frequency (how often the item was seen), recency (how recently the item was seen), and spacing (the gap between item presentations). Based upon these results, Anderson and Schooler (1991) have developed a generalized mathematical model of human memory. Furthermore, as will be discussed, they argue that human memory has adapted to a structure that exists in the environment [Anderson and Schooler, 1991].

Similarly, Burrell (1980) has analyzed the lending patterns of books in large libraries. This work has resulted in a stochastic model for predicting patterns of book borrowed from a library collection. The only data required by the model are the size of the collection and the number of times each book is loaned during specified periods of time. Note that, unlike in the memory approach, frequency is the only predictor variable. However, in the case of

libraries, only books that are not currently checked-out are available for borrowing. Thus, frequency, recency, and spacing are likely to be highly correlated. This does not hold for digital libraries, where documents may be retrieved concurrently by many users.

It was our expectation that these models may apply to large repositories having high access rates. Specifically, we wished to determine if recency and frequency rates of past documents access would accurately predict future document requests¹. To examine this idea, we used the memory model of Anderson and Schooler (1991) to predict document access in a large, multimedia repository.

The repository selected was the Georgia Institute of Technology World Wide Web (WWW) database² [Berners-Lee et al., 1994]. At the time of our analysis, this repository contained over 2000 multimedia documents, and received as many as 12,000 documents accessed per day, with some documents accessed over 4000 times per week.

In the next section we describe the model for computing and predicting document request rates, based on frequency and recency rates of past document accesses. We describe the test dataset and show results of applying the model. We then describe extensions to the basic approach. First, we combine the recency and frequency analyses. Second, we incorporate repository structure and document type. We conclude with ways in which our prediction accuracy might be improved, and sketch applications in the areas of design of information systems and interfaces, and their caching algorithms.

The Model

Human memory has a long tradition of research in the psychology literature. One focus of this research is on the relationship of the time delay between when an item is presented and subsequent performance on recall. A related focus is on the number of practice trials for items and subsequent performance on recall. As might be expected, the results show that

¹For the purpose of this paper, the terms “access,” “request,” and “retrieval” are synonymous

²The home page Universal Resource Locator (URL) is <http://www.gatech.edu/TechHome.html>.

shorter delays and higher amounts of practice lead to better recall performance.

Based on a review of this literature, Anderson and Schooler (1991) argue that the relationship between the time when an item is first presented and subsequent performance (retention) is a power function. Therefore, under a logarithmic transform, a linear relationship is found between the time measure and performance measure. Similarly, they argue that the relationship between trials of practice and performance is also a power function. This *power law of practice* is a robust finding in the literature, and is also exhibited by some artificial intelligence (AI) learning models (e.g., knowledge compilation and chunking models of learning [Anderson, 1987, Newell and Rosenbloom, 1981]).

Interestingly, Anderson and Schooler (1991) propose an environmental explanation for the existence of these relationships. They suggest that, in fact, the memory system has adapted to the structure of the environment. The memory system is estimating the likelihood that particular memory items will be needed and, through adaptation, is trying to make the high likelihood items available. In essence, they argue that the memory system infers the probability that a particular memory item will be needed at the present moment. This, they call the *need probability*, or p .

If one spends a quick moment considering these need probabilities, it becomes evident that most of the memory items are not needed *now*, and a few are needed frequently. Thus, the distribution of need probabilities shows a large number of cases near zero, with a tail of a few need probabilities closer to 1. Because of this J-shaped distribution, Anderson and Schooler argue that it is more convenient to think about this distribution in terms of “Need Odds.” This term is simply $p/(1-p)$. This transformed distribution will vary from 0 to infinity, instead of from 0 to 1. Furthermore, the logarithmic transform of the “Need Odds” will vary from minus infinity to infinity. Note also that this transformation makes the many types of data more amenable to statistical analysis, since there are no upper and lower bounds.

In order to determine how past usage of information predicts future usage, Anderson and Schooler (1991) developed an algorithm for analyzing the occurrence of items in large repositories. They applied their algorithm to the analysis of repositories of information in terms of frequency, recency, and spacing rates of items in these sources. In particular, they analyzed three repositories: (1) the occurrence of words in newspaper headlines from the *New York Times*, (2) utterances made to children extracted from the CHILDES database, and (3) the electronic mail addresses from mail sent to one person.

In the next sections, we describe our test dataset, and show how we applied their model and algorithm to compute and analyze document access rates in the Georgia Tech WWW database.

The Dataset

The dataset used in our analysis was the log file of accesses to the Georgia Tech WWW repository. This repository is part of the Internet accessible World Wide Web [Berners-Lee et al., 1992, Berners-Lee et al., 1994]. The WWW protocol provides seamless hypermedia access to the widely distributed and heterogeneous collections of existing information resources, including but not limited to: Archie [Emtage and Deutsch, 1992], Gopher [Alberti et al., 1992], File Transfer Protocol (FTP) [Postel and Reynolds, 1985], Network News Transfer Protocol (NNTP) [Kantor and Lapsley, 1986], and Wide Area Information System (WAIS) [Addyman, 1993]. In addition, WWW access logs are particularly interesting for our analysis because Merit estimates of NSFNET backbone traffic show that WWW is one of the fastest growing information resources³.

Our analysis used the log file of accesses during a three month period, January 1 through March 31, 1994⁴. From the log file, we removed all accesses made by Georgia Tech machines. We feel that these accesses add noise to the data because they often represent users testing

³See <http://www.cc.gatech.edu/gvu/stats/NSF/merit.html>.

⁴Due to a system failure, data from three days in March are absent.

new documents or default document accesses made by client programs.

The trimmed log file comprised 35 megabytes of data, with a mean record length of 100 bytes and totaling roughly 305,000 requests. The number of requests ranged from 300 to 12,000 document per day, with a mean of 3379 accesses per day over the three month period. Some individual documents were accessed up to 4000 times per week.

The Repository: A Dynamic Information Ecology

The Georgia Tech WWW repository epitomizes what we call a *dynamic information ecology*. Because it is part of a world-wide, distributed, multimedia information network, the number of documents, their links, and structure are constantly in flux. For example, the repository can change internally, as information is supplied by over 36 information providers. The repository can also change externally, as information providers elsewhere on the Internet add and change pointers to documents located at Georgia Tech.

At the time of our analysis, the WWW repository contained over 2000 multimedia documents. Documents used several forms of media, including text, postscript, GIF, jpeg, mpeg, audio, CGI script requests ⁵, etc.

The repository exhibits irregularities that would pose problems for conventional indexing methods. For example, accesses to the database are subject to fluctuations, which result, in part, from reduced weekend activity. However, the repository also contains many temporally dependent documents (e.g., documents containing the monthly updates of NSFNET backbone statistics). In these cases, the content of documents may change, while the document names do not. As expected, these temporally dependent documents are accessed with high frequency and recency during a few days of the month (i.e., the access rates spike), and then decrease gradually.

There are many other types of irregularities in the way documents are accessed. These

⁵CGI stands for Common Gateway Interface. CGI causes scripts to be executed by the servers, rather than document retrieval.

include, for example, client-side caching of documents and images, CGI script requests (where each request may result in a unique document request), non-graphical based clients (which do not request images and movies embedded in documents), document insertions, deletions and renamings. Finally, the number of documents and their links continues to change and grow. For the repository being studied, the document space grew on the average of 2.3% per day.

It is our belief that these irregularities, which exist both in the dynamic nature of document structure and in their requests, are fundamental characteristics of dynamic information ecologies. Furthermore, it seems likely that repositories with these characteristics will become increasingly common. It therefore becomes important to develop robust techniques and methods for understanding and predicting information seeking patterns in such ecologies.

Frequency Analysis

In our analysis, we were interested in determining the relationship between the number of document requests during a period (called the *window*) and the probability of access on a subsequent day (called the *pane*). This analysis can be viewed as a parallel to the practice function in human memory research. In this case, given the frequency of past document requests, we are interested in determining the probability of new requests.

Following the algorithm described in Anderson and Schooler (1991), we computed the frequency of document access during each 7-day window in the dataset, and measured their probability of access during the next day (i.e., day 8). We selected a window of 7 days because we intuitively felt that this window would encompass the typical fluctuations inherent in the calendar week.

We illustrate how the frequencies and probabilities are calculated with the following example. During Window 1 (day 1 through 7), we find that documents A and B are accessed 6 times. We then find on Pane 1 (day 8) that A is accessed but B is not. Therefore, for the

first window and pane, the probability of access for the frequency of value 6 is the sum of accesses in the pane (1+0) divided by the number of accesses in the window (1+1), or .50. These probabilities are calculated for all frequency values.

Continuing our example, we find in Window 2 (day 2 through 8) that documents C and D are accessed 6 times. We find on Pane 2 (day 9) that neither document is accessed. Our new probability of access for the frequency of value 6 is the sum of accesses in the two panes (1+0+0+0) divided by the number of accesses in the two windows (1+1+1+1), or .25. In this way, probabilities are computed for all frequency values, for all windows and panes in the dataset.

Results

Using the algorithm described above, a window size of 7 days, and a pane size of 1 day, we computed the probabilities of access all documents based on frequency of access during each previous window in the dataset.

Figure 1 shows the relationship between the frequency of document accesses during the past seven days (for frequencies < 100) and the probability that it will be accessed on the current day. As can be seen, the plot shows a power relationship and is similar to those typically shown by practice plots found in the memory literature. In short, as the frequency of document accesses increases, the probability of access on the current day increases as a power function.

Insert Figures 1 and 2 about here

The frequency analysis is complex due to the sheer volume of accesses made to the repository. We found that a small number of documents are requested with very high frequency values within windows. Thus, the probability of access within a pane is based on a very small

number of documents, and hence subject to large variability. To account for this effect, we combined frequency values into clusters. Values from 1 to 100 were unchanged, values from 100 to 1000 were rounded to the nearest hundred, and values from 1000 to 4000 were rounded to the nearest 1000⁶.

Figure 2 shows a logarithmic transform between the frequency of document cluster accesses during the past seven days and the probability that it will be accessed on the current day. The plot shows a strong power relationship between frequency and probability of access. This relationship again mirrors the power law of practice found in memory research.

As proposed by Anderson and Schooler (1991), a more interesting comparison is to plot the log of frequencies of access against “log(Need Odds).” Recall that if p is the probability then

$$\text{Need Odds} = p/(1-p)$$

Figure 3 plots the relationship between need probability and frequency, with a logarithmic transform (for clustered frequencies). As can be seen, there is a strong relationship between “log(Need Odds)” and the log of the frequencies, $F(1, 94) = 246.53$; $p < .001$; $MS_E = 124.32$; $R^2 = .72$. The regression equation:

$$\text{Log(Need Odds)} = .99 \text{ Log(Frequency)} - 1.30$$

accounts for 72% of the variance. The regression coefficient (.99) shows that the regression equation slightly underestimates actual probability of access.

Recency Analysis

We also analyzed the recency of accesses for documents during a 7-day window and measured their probability of access in a 1-day pane (day 8). This analysis parallels the retention

⁶Note that these clusters are log derivatives.

function in human memory research. In this case, we are looking at the probability of document access on the eighth day (the pane) based on how many days have elapsed since the document was last requested in the window.

Recency probabilities are computed like the frequency probabilities. We illustrate the computation with the following example. During Window 1 (day 1 through 7), we find that documents A and B were most recently accessed two days in the past from Pane 1 (e.g., day 6). We then find on Pane 1 (day 8) that A was accessed but B was not. Thus, for recencies of value 2, the probability of access is .50 $((1+0)/2)$. Continuing, we find in the next window, Window 2 (day 2 through 8), that documents C and D were most recently accessed two days in the past. We find on Pane 2 (day 9) that both documents were accessed. The new probability of access for recency of value 2 is now .75 $((1+0+1+1)/4)$.

Results

Figure 4 plots the probability of access on day 8 against how many days have passed since the document was last accessed in the previous 7-day window⁷. The plot shows the steep negative slope typically found in retention plots in memory research. As days since the last access elapse, the document is much less likely to be accessed on day 8.

Insert Figure 4 about here

Figure 5 plots the relationship between need probability on day 8 and recency of document access, with a logarithmic transform. As can be seen, there is a strong relationship between “log(Need Odds)” and “log(days)”, $F(1, 5) = 56.17$; $p = .001$; $MS_E = 3.71$; $R^2 = .92$. The regression equation is

⁷We have defined Day 1 to mean the document was accessed one day previous to the pane (or the last day of the window), in contrast to Anderson and Schooler (1991) who used Day 0.

Recency Day	Mean	Range	S.D.
One	411.1	555.0	126.5
Two	133.5	274.0	59.7
Three	87.4	220.0	46.3
Four	65.6	140.0	34.4
Five	55.0	154.0	31.0
Six	46.5	93.0	24.0
Seven	46.4	97.0	25.4

Table 1: Mean, Range, and Standard Deviations for document frequency on each recency day in the 7-day window.

$$\text{Log(Need Odds)} = -1.15 \text{ Log(Days)} + .41$$

and accounts for 92% of the variance. Again, the regression equation slightly underestimates actual probability of access.

Insert Figure 5 about here

In summary, the recency analysis showed the logarithmic relationship typically found in the retention memory literature. In addition, recency proved to be a much better predictor than frequency.

Combining Recency and Frequency

Despite these findings, the relationship between recency and frequency in predicting document access remained unclear. To examine this relationship, we analyzed the recency data to determine the number of document accesses that occurred for each day in the recency windows. We then computed the mean number of documents for each day for all sampled windows. Table 1 shows the mean, range, and standard deviations for each recency day. For each day, the range and standard deviations confirm the dynamic nature of the dataset,

which, as reported earlier, showed approximately a 2% growth in the document space per day.

Figure 6 shows the 3-D plot of the mean number of documents accessed for each recency day for each window during the month of March (the plot for the full data set is not displayed due to its large size). The plot provides a visualization that reveals a kind of cascading effect. As predicted by the model, when a large (or low) number of documents occurs with a recency of one day, similar effects are observed six days later for the number of documents occurring with a recency of seven days.

Insert Figure 6 about here

We next analyzed whether the mean number of documents also followed a power law relation with respect to: (1) each day, (2) $\log(\text{Need Odds})$ and (3) $\log(\text{days})$. Figures 7, 8, and 9 show the results of these analyses. The regression equations for the latter two are:

$$\text{Log}(\text{Need Odds}) = 1.04 \text{ Log}(\text{Days}) - 5.60$$

$$\text{Log}(\text{Mean Number of Docs}) = -1.12 \text{ Log}(\text{Days}) + 5.83$$

We observe a very strong fit for the logarithmic transformations, with both regression lines accounting for over 96% of the variability. Note that as a direct result of the robust fit for recency, the regression lines for the mean number of documents with respect to “Need Odds” and days are essentially the same, except for the direction of the slopes. Also, we note that the changes in the number of document accesses decreases through time across days, following a power law relationship (see Figure 7).

Insert Figures 7, 8, 9 about here

<i>Recency Days</i>	All Links	Text Links	Image Links	In	Out
One	3.21	2.73	0.16	2.99	2.86
Two	1.99	1.67	0.08	2.19	1.72
Three	1.42	1.22	0.04	1.82	1.24
Four	1.28	1.06	0.03	1.68	1.08
Five	1.20	1.06	0.02	1.71	1.06
Six	1.04	0.94	0.02	1.90	0.94
Seven	1.63	1.39	0.04	1.81	0.37
Recency r	-.71	-.77	-.80	-.71	-.69

Table 2: Mean number of links, text links, image links, relative in centrality, and relative out centrality for documents per recency day. The last row shows the Pearson correlation between days since last access and links.

Repository Structure and Document Type

Two fundamental hallmarks characterize hypermedia information repositories. The first is their interlinked document structure. The second is their ability to display documents of many types (e.g., text, video, audio, images). In this section, we examine the relationship between these attributes and document recency and frequency rates.

Repository Structure

Document interconnectivity was measured by counting the number and type of links per document. Links to other documents were classified into one of 6 types: links to (1) audio, (2) video, (3) text, (4) images, and (5) other. In addition, we calculated the relative *in centrality* and *out centrality* for each node in the repository [Rivlin et al., 1994]. These metrics are defined in terms of the length of the path from any arbitrary node to another. A node that can easily access other nodes is defined to have a *high relative out centrality*. A node that is easily accessed by other nodes is defined to have a *high relative in centrality*.

Table 2 shows the mean number of overall links, links to text documents, links to images, and relative in and out centrality for documents in each day of the window. The last row shows the Pearson correlation between recency of access and link number and type. The

	All Links	Text Links	Image Links	In	Out
Frequency r	-.22	-.18	-.21	-.23	-.18

Table 3: Pearson correlation between frequency of access and link number, type, and in and out degrees.

	Text	Images	Audio	Video
Recency r	-.77	-.79	-.44	-.15
Frequency r	-.15	-.19	-.10	-.09

Table 4: Pearson correlation between frequency and recency of access with document type.

high correlations suggest that as documents age, the mean number of links per document and their relative in and out degree decreases. In short, recently accessed documents have a higher degree of interconnectivity

Table 3 shows the correlation between frequency of access and link number and type. While the correlations are negative, they are much weaker than the recency results.

Document Type

Additionally, we classified requested documents in the database into one of 5 types: (1) text, (2) images, (3) sound, (4) video, and (5) other. Table 4 shows the correlation between recency and frequency of access and document type. The two most common document types are text and image. The results again show that correlations are much higher for the recency measure. As documents are more recently accessed, they are more likely to be text and image documents. The correlations between document frequency and document type are much weaker.

Discussion

In this paper, we analyzed patterns of document requests in a non-standardized, heterogeneous, and inherently chaotic repository, or what we have called a dynamic information ecology. The results indicate that recency and frequency of past document access are strong

predictors of future document access. We note that these results were obtained without knowledge of the *contents* of the documents. Moreover, our results come from analyzing the access patterns of real users, with real information needs – not users in experimental or laboratory studies.

In our analysis, recency proved to be a stronger predictor than frequency. In addition, better predictions were obtained when frequency and recency were combined. Analyses of document link structure and document type incorporated with recency and frequency rates revealed stronger relationships with recency. The results suggested that more recently accessed documents were more likely to have a greater number of links. They were also more likely to be text and image documents.

As a caveat, we note that the strength of our predictions may be an artifact of the window size used in our analyses. Moreover, while the structural analysis was based on overall document interconnectivity, it did not take into account the extent to which recently accessed documents were linked *to one and another*. The latter analysis would establish a kind of *context* effect for document access.

In future work, we plan several additional analyses. First, we plan to conduct analyses in which window and pane sizes are manipulable parameters. We will also keep track of the context of document interconnectivity. Second, we hope to perform these analyses using access data from other repositories. Third, we plan to incorporate an analysis of spacing effects. Research on human memory has found relationships between the time delay between item presentations and performance on recall. Similarly, repositories contain documents that are requested with different time lags. For example, some documents may be requested daily (e.g., the weather) or monthly (e.g., monthly statistics), while others may display less consistent time lags. We expect that these additional analyses will lead to better predictions, as well as equations and parameters that can be generalized to databases of different sizes and exhibiting different access rates.

Applications

These results have several implications for the design of information systems that exhibit high access rates. First, information retrieval systems should support users with particular information needs in efficiently finding desired documents. Frequently, information access is performed through the use of indexes. However, indexing textual material is difficult and time consuming, often resulting in searches occurring under weak indexing conditions. Effective indexing of non-textual material is even more complex. Furthermore, the indexing vocabulary may be inappropriate or difficult for the user to employ. For example, analysis of document and book retrieval using computer retrieval techniques show that typically searchers find 10-40% of the relevant materials while getting back 30-60% of unwanted items [Landauer et al., 1993]. Therefore, we seek methods that may enhance weak, under-specified, or inefficient indexing schemes. Our results suggest that past recency and frequency data can be used to determine optimal ordering of retrieved items.

Second, the results can be used to inform design decisions and compare among alternative designs of information systems. For example, a *Cost-of-Knowledge Characteristic* function has been proposed for characterizing information access in dynamic displays [Card et al., 1993]. An important aspect of this function is the formulation of a probability density function that weights items according to their values. These items could be documents accessed in an information system or interface actions in a GUI. It seems an easy extension to use results from the present approach to derive probability density functions for different tasks and systems.

Third, frequency and recency data can be used to help suggest navigation strategies for users of large hypermedia systems [Rivlin et al., 1994]. Fourth, access log visualization software [Pitkow and Bharat, 1994] can be augmented to display deviations from predicted values on a per document basis for any given time frame. This use would enable database designers and maintainers to maximize the efficiency of their document space. For example,

the model's predictions could support decisions involving the addition of hyperlinks to under-accessed but important documents, as well as the removal of unnecessary documents from the document space.

Finally, the most natural application of these analyses lies in the design of caching algorithms used by multimedia systems. Our results suggest strategies for determining what items should be stored in the cache, and for what duration. While it is doubtful that the results will greatly effect server side caching, large scale caches, often called proxies, may benefit⁸.

For example, the HTTP proxy/server distributed by CERN [Luotonen and Atlis, 1994] allows for the administrator of the proxy to hard-code key variables. These variables include: 1) the maximum time to keep used and unused cached files, 2) the time the cached files were last modified and 3) when and how often to perform garbage collection. We feel that more adaptive and access-based strategies can be implemented that are based on recency analyses. Such strategies would remove the need for heuristics by optimizing the cached items based on long-term retrieval patterns. In addition, the caching strategies can adjust to dynamic access patterns.

Additional analyses of the data help to formulate the caching strategy. Figure 10 plots the predicted number of documents that would be in the cache and requested on the target day (day 8). This plot essentially combines Figures 4 and 7. The Y-axis of this graph represents the mean number of documents occurring on each day multiplied by the probability of the document being requested on the target day. The rapid drop off in the number of documents needed in the cache suggests that decreasing the size of the sampling window might lead to a more gradual drop off. We have yet to examine this possibility.

Figure 11 plots the cumulative hit rate⁹ from Figure 10, as a function of the number

⁸Caching at the server side level is primarily a hardware and operating systems issue, whereas larger scale caching (e.g., caching documents to avoid satellite/trans-ocean communication costs) is typically constrained by the amount of available memory and efficient document removal (garbage collection) algorithms.

⁹Hit rates can be defined as the number of documents in the cache divided by the number of requested

of days documents are stored in the cache. Again, we note the presence of the power law relation in examining the predicted hit rates for each day. For example, from this figure, we see that if the cache keeps documents for up to four days since their last access, the cache will contain the requested item on the next day (day 8) 51.5% of the time. Similarly, on day seven, if all requested documents were cached during the sampling window, there is only a 46% chance of a document in the cache being requested on the target day.

Insert Figures 10, 11, 12 about here

Figure 10 and 11 thus strongly suggest that the most important data point for a caching algorithm is the pool of documents accessed with a recency of one day (or, the last day in the window). With this in mind, Figure 12 shows the hit rate of documents stored in the cache with a recency of one day as a function of all windows (from January 1 through March 31). This provides us with a window by window analysis of the model's predictive abilities.

A closer inspection of this plot reveals a support line of roughly 44%, a ceiling of 84%, with a mean of 67% (see Figure 12). Finally, we note that the large drops in Figure 12 occur mostly on weekends. We believe that this results from the presence of documents requested on Fridays that were not requested on the following Saturdays. Thus, the observed drops seem natural when taking into account the decreased usage of computer resources on weekends. Furthermore, this suggests that the overall predictive abilities of the model might be increased by excluding weekends from the analysis, thus pushing the support, mean, and ceiling lines up higher, though analysis has yet to be conducted.

The above results and analyses provide the motivation for a simple, least-recently used (LRU)[Tanenbaum, 1992], garbage collection algorithm. A sketch of the algorithm is presented in Figure 13. Note that the overall caching algorithm that provides users with the documents. For our purposes, we define hit rates as the mean number of documents per day of recency multiplied by the probability of the document being needed on the target day.

most current version of documents may require that all incoming requests be checked to determine if the cache still contains the most current version of the requested document. Nonetheless, the proposed algorithm would still enable correct caching of roughly 2/3 of the incoming requests. This hit rate compares quite favorably with existing caching/proxy systems [Glassman, 1994, Smith, 1994].

Insert Figure 13 about here

Memory and Document Access

In closing, we are intrigued by the similarities between our results on document access and the original claims from the memory literature. Recall that Anderson and Schooler (1991) were arguing that the human memory system has adapted to the structure and demands of its environment. Our concern lay in predicting document access, by relying on previous access rates. From the perspective of Anderson and Schooler (1991), we find that human memory mirrors the observed behavior of environmental events. In our work, we notice access patterns that appear to mirror the behavior of human memory. These strong parallels suggest a kind of group memory that functions in a dynamic information ecology in ways similar to the memory of individuals.

We also note that the repositories studied by Anderson and Schooler (1991) originated in humans. That is, in their datasets, humans are the ones producing the newspaper headlines, talking to their infant children, and sending e-mail. This is also true of WWW databases. It would be interesting to see if these predictions extend to events in natural – rather than human-created – environments.

Acknowledgments

M. Recker is supported by the Army Research Institute for the Behavioral and Social Sciences under Contract No. MDA-903-90-K-112. We thank Jim Foley, Mark Guzdial, Scott Hudson, Tom Landauer, Peter Polson, and Peter Pirolli for comments on an earlier version of the paper. The programs (in C) and scripts (for Unix) used in performing the analyses described in this paper are available for anonymous FTP from: `ftp.cc.gatech.edu` in `/pub/gvu/www/pitkow/prediction_analyzer/prediction_analyzer.tar.Z`.

References

- [Addyman, 1993] Addyman, T. (1993). WAIS: Strengths, Weaknesses, and Opportunities. In *Proceedings of Information Networking 93 Conference*, London. Meckler.
- [Alberti et al., 1992] Alberti, R., Anklesaria, F., Lindner, P., McCahill, M., and Torrey, D. (1992). The Internet Gopher Protocol: a distributed document search and retrieval protocol
- [Anderson, 1987] Anderson, J. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review*, 94(2):192–210.
- [Anderson and Schooler, 1991] Anderson, J. and Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science*, 2(6):396–408.
- [Berners-Lee et al., 1992] Berners-Lee, T., Cailliau, R., Groff, J., and Pollermann, B. (1992). World-Wide Web: The information universe. *Electronic Networking: Research, Applications, and Policy*, 1(2):52–58.
- [Berners-Lee et al., 1994] Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H., and Secret, A. (1994). The World-Wide Web. *Communications of the ACM*, 37(8):76–82.

- [Burrell, 1980] Burrell, Q. (1980). A simple stochastic model for library loans. *Journal of Documentation*, 36(2):115–132.
- [Card et al., 1993] Card, S., Pirolli, P., and Mackinlay, J. (1993). The cost-of-knowledge characteristic function: Display evaluation for direct-walk dynamic information visualizations. In *CHI'94: Human Factors in Computing Systems*, pages 238–244, Boston, MA. ACM.
- [Emtage and Deutsch, 1992] Emtage, A., and Deutsch, P., (1992). Archie: an electronic directory service for the Internet. In *Proceedings of the Winter 1992 Usenix Conference*.
- [Glassman, 1994] Glassman, D. (1994). A Caching Relay for the World Wide Web. In *Proceedings of the First International World Wide Web Conference*, Amsterdam. Elsevier.
- [Kantor and Lapsley, 1986] Kantor B., and Lapsley, P. (1986). Network News Transfer Protocol - a proposed standard for the stream-based transmission of news. *Internet Request for Comments, RFC 977*.
- [Landauer et al., 1993] Landauer, T., Egan, D., Remde, J., Lesk, M., Lochbaum, C., and Ketchum, D. (1993). Enhancing the usability of text through computer delivery and formative evaluation: the superbook project. In McKnight, C., Dillon, A., and Richardson, J., editors, *Hypertext: A psychological perspective*. Ellis Horwood, New York.
- [Luotonen and Atlis, 1994] Luotonen, A. and Atlis, K. (1994). World-Wide Web Proxies. In *Proceedings of the First International World Wide Web Conference*, Amsterdam. Elsevier.
- [Newell and Rosenbloom, 1981] Newell, A. and Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In Anderson, J., editor, *Cognitive skills and their acquisition*. Lawrence Erlbaum, Hillsdale, NJ.
- [Nielsen, 1990] Nielsen, J. (1990). *Hypertext & Hypermedia*. Academic Press, San Diego, CA.

- [Pitkow and Bharat, 1994] Pitkow, J. and Bharat, K. (1994). WEBVIZ: A Tool for World Wide Web Access Log Visualization. In *Proceedings of the First International World Wide Web Conference*, Amsterdam. Elsevier.
- [Postel and Reynolds, 1985] Postel, J., Reynolds, J. (1985). File Transfer Protocol (FTP). *Internet Request for Comments, RFC 959*.
- [Rivlin et al., 1994] Rivlin, E., Botafogo, R., and Shneiderman, B. (1994). Navigating in hyperspace: Designing a structure-based toolbox. *Communications of the ACM*, 37(2):87–96.
- [Smith, 1994] Smith, N. (1994). What can Archives offer the World Wide Web? In *Proceedings of the First International World Wide Web Conference*, Amsterdam. Elsevier.
- [Tanenbaum, 1992] Tanenbaum, A. (1992). *Modern Operating Systems*. Prentice-Hall, New Jersey.